CSE 676 Deep Learning Spring 2020

Image Captioning with Transformer

Srikar Challa
50312357
University at Buffalo
Buffalo, NY 14260
srikarch@buffalo.edu

Srisai Karthik Neelamraju
50316785
University at Buffalo
Buffalo, NY 14260
neelamra@buffalo.edu

Anantha Srinath Sedimbi
50315869
University at Buffalo
Buffalo, NY 14260
asedimbi@buffalo.edu

Abstract

In this paper, we try to define a new model for improving performance of image captioning. The task of image captioning is to automatically describe the contents of any naturally photographed image, understanding the visual scene presented by the image. This is one of the fundamental problems that involves multiple branches of Artificial Intelligence (AI) such as Computer Vision, Language Modeling, and Natural Language Processing. It has a variety of applications, such as aiding visually impaired and automatic image indexing. Vinyals et al. in [1] make use of a convolutional neural network (CNN) combined with deep recurrent neural network (RNN) architecture and have achieved good results on MSCOCO [16] and Flickr30k [17] datasets. Yet they assert that there is an immense scope to improve. Xu et al. in [2] managed to generate much more meaningful captions by employing attention mechanism along with a sequential model for text generation such as Long Short-Term Memory (LSTM) networks. Vaswani et al. in [18] developed a novel architecture, named Transformer, that is based solely on attention mechanisms to achieve machine translation tasks. In this work, we propose using the transformer model to generate image captions. We employ the Xception [19] deep CNN architecture to extract features from an image and feed these features as the input to the transformer which then generates text as the model output.

1 Introduction

Recognizing the contextual relation between entities of an image involves multiple challenges such as accurate object detection, categorization of objects, recognizing attributes of the scene and then providing a semantic description of these related elements as a whole. Unlike regular image-classification problems, image captioning does not make use of iconic images [16] (images that are clear of any clutter, occlusion, background, with the subject almost in the center) as inputs. In fact, it uses natural images that contain many objects in one image and perhaps no specific subject in the center with fair importance to the background of the scene. Due to ambiguity of object sizes, orientation and count, regular image classification models do not perform well on such images. Moreover, contextual reasoning between the objects has to be learned by the model from training captions itself, making it a much harder problem. Powerful computer vision models help to address the former problem of object identification. Many state of the art pretrained convolutional nets such as VGG-16 and VGG-19 [10] can be employed with less difficulty. However, such models are parameter heavy. For example, VGG-16 which has one of the best classification accuracies on ImageNet [7] has about 140 million parameters. An improvement in the feature extraction can be to reduce parameter size while having similar detection capabilities. InceptionV3 has good accuracy with reduced parameters – it makes use of smaller convolutional kernels by bringing down the total no of trainable weights in the model. But considerable improvement in caption generation was achieved by Xu et al. [2] mainly from the attention. Attention allows salient features to dynamically come to the forefront as needed. Rather than explicitly employing an attention formulation such as the one used by Bahdanau, et al. [4] into a sequential decoder (such as RNN, LSTM) we make use

of a Transformer [18] decoder which implicitly allows for parallel multi-head attention. Transformers have proven capability to outperform recurrence-based models in sequence to sequence generation tasks. Though image captioning is different to more regular applications of the Transformers such as sentence translation from one language to another, we see a possibility to train a transformer decoder to map embedded feature vectors to meaningful word sequences in one language itself (here, English). Contributions of this project are:

- We study the performance of computationally efficient feature extractors such as Xception and sequential text generators such as Gated Recurrent Unit.
- We study the usability and efficiency of Transformer as a text generator instead of sequential models.
- Further, we present cases where the BLEU evaluation metric tends to be ambiguous.

2 Related Work

Several methods have been proposed in the recent years for addressing the task of image caption generation. Most of them have been inspired by the advances in generating text from neural networks for machine translation [3, 4, 5]. These models typically follow a CNN and an RNN based combined network that generates captions. One of the most impactful works in this area was proposed by Vinyals, et al. in [1], where they use GoogLeNet [6] network pre-trained on ImageNet [7] for extracting features from an image and then use an RNN with LSTM [27] units for generating text from the extracted features. Similar works were proposed by Mao, et al. in [8] and Kiros, et al. in [9]. Xu, et al. in [2] have proposed an attention-based mechanism for image captioning, where they use VGG-16 [10] network architecture for feature extraction and use soft/hard attention [4] for training the RNN. Many other works followed a similar approach to image captioning and were extensively detailed by Hossain, et al. in [11]. More recently, Lu, et al. in [12] and Qi, et al. in [13] have proposed extensions to the BERT [14] architecture for developing a vision-language model for learning joint representations of image content and natural language. The model in [13], which they named ImageBERT and pre-trained on the Conceptual Captions Dataset [15], has achieved state-of-the-art results on both MSCOCO [16] and Flickr30k [17] datasets. Vaswani, et al. in [18] proposed an architecture, called Transformer, that is based solely on attention mechanism. They successfully replace RNNs for machine translation tasks using this model, while gaining significant improvements over the regular models. In this work, we propose employing the Xception [19] network pre-trained on ImageNet [7] for extracting image features and passing these features to a Transformer [18] model that involves multiple attention heads for generating captions.

3 Model Architecture

3.1 Model One (M1)

The base model in [1] uses a GoogLeNet [6] network pre-trained on ImageNet [7] for extracting features from an image and then employs an RNN with LSTM [27] units for generating text from the extracted features. The authors in [2] make use of a similar architecture, but replace the GoogLeNet network with VGG-16 [10]. However, the VGG-16 architecture has very large number of parameters (nearly 140 million) as described in Table 1. This means difficulty in deploying the network and a longer inference time. Several other CNN architectures are proven to be much superior to VGG-16 both in terms of reducing the computational cost as well as providing better performance on ImageNet dataset. One of these newer models is the InceptionV3 [22] network, which builds upon the inception architecture of the GoogLeNet. This model has only about 24M parameters and has obtained a much better top-1 accuracy on ImageNet. Since the ImageNet dataset contains many common real-world objects, a model pre-trained on it generally provides good performance on smaller datasets like MSCOCO [16], Flick30K [17] and Flickr8K. Hence, for our first model, we employ an InceptionV3 feature extractor instead of GoogLeNet.

Model	Top-1 Accuracy	Top-5 Accuracy	Parameters	Depth	Weights Size
VGG-16 [10]	0.713	0.901	138,357,544	23	528MB
VGG-19 [10]	0.713	0.900	143,667,240	29	549MB
InceptionV3 [22]	0.779	0.937	23,851,784	159	92MB
Xception [19]	0.790	0.945	22,910,480	126	88MB

Table 1: Comparing performance of different CNN architectures on ImageNet

We also make a slight modification in the decoder architecture when compared to the original papers [1, 2]. Gated Recurrent Units (GRUs) were proposed by Cho, et al. [26] as an alternative to LSTM [27] units. They have fewer gates than LSTM and are relatively faster to train. Therefore, for our first model, we replace the LSTM-based decoder with a GRU-based decoder. We also employ the soft attention mechanism that was implemented in [2], so that the model can automatically learn to fix its concentration on certain parts of the image that it should be focusing on when generating corresponding caption word.

3.2 Model Two (M2)

For our second model, we only replace the InceptionV3 [22] architecture in the first model with Xception [19] network and keep the decoder part of the network unchanged. Xception model was inspired from the former model and replaces the inception modules in that model with depth-wise separable convolutions. This reduces the number of parameters in the network by about 1 million. Xception model achieves a slight improvement on the original model, while requiring a lesser the number of network parameters. The authors suggest that this performance gain is due to the efficient use of the model parameters. Both the models have same similar input and output shapes. Therefore, in sum, our second model M2 can be described as employing an Xception feature extractor, along with a soft-attention [4] based RNN-decoder with GRUs [26].

3.3 Model Three (M3)

Despite their decent performance in generating the image captions, the previous two models still make use of RNN-based architectures. Even though these models are powerful and vital for modeling sequential data, they process the data in a sequential manner. RNNs suffer from the vanishing gradient problem and it is really hard for them to learn long-range dependencies. These drawbacks in turn inhibit the learning speed and the performance of the model. Also, if we use an RNN-based architecture for caption generation, attention has to be implemented explicitly as a separate entity.

We propose a major modification of using a transformer [18] model instead of recurrent architectures for the text generation task. This novel model was based solely on self-attention mechanism and allows multi-head attention. According to the authors, multi-head attention is proven to learn "multiple tasks", meaning that some attention heads have been observed to recognize start of a sentence while some other focused on the end for example. This could benefit the image caption generation task a lot.

Despite these advantages, a challenge in the transformer architecture is the difficulty in detecting temporal correlation in the input sequences (which a sequential model such as RNN or LSTM can do by default). It sees input sequence like a bag of words unless we make use of positional encoding, a fact which perhaps hindered adaptation of Transformers into the task of image captioning leading to a number of papers that repeatedly make use of RNN or LSTM or an equivalent model. Hence, we make use of positional encoding along with word embedding similar to Vaswani et al [18].

Table-2 summarizes the architectures of the three models that have been implemented.

Model	Architecture
M1	InceptionV3 feature extraction + RNN with GRUs + Soft Attention
M2	Xception feature extraction + RNN with GRUs + Soft Attention
M3	InceptionV3 feature extraction + Transformer Model

Table 2: Summary of the three models implemented in this project

4 Experiments

4.1 Evaluation Metric

Given the ground truths, several metrics are known for evaluating the quality of machine-generated text. BLEU score [20] is the most commonly reported metric in the area of image captioning. Hence, we use the BLEU score to evaluate the performance of our models. BLEU stands for BiLingual Evaluation Understudy and it is an algorithm that relies on the central idea of a machine translation being good if it is close to a professional human translation. It computes similarity between predicted text and reference text using a form of modified precision. It is a number between 0 and 1; the higher the score, the better it is. BLEU scores are typically evaluated by considering groups of *n* words or word *n*-grams. BLEU-*n* corresponds to the modified precision evaluated using *n*-grams. Here, we report the BLEU-1, 2, 3 and 4 scores for our models.

4.2 Implementation

We train our models on the Flickr8k dataset which has 8,091 images and 5 reference captions associated with each image. For our experiments, we limit the vocabulary size to 5,000. As the authors of [2] note, there is a difference in the dataset splits across different algorithms [1, 2, 21] due to a lack of standardized splits. Nevertheless, they also advocate that differences in splits does not bring about a significant change in the overall performance. Consequently, we use our own split (80% training data) for evaluating our models. Every image is reshaped to $299 \times 299 \times 3$ in order to comply with the input shape of the InceptionV3 [22] and the Xception [19] networks. The reshaped image is then propagated through one of these two pre-trained networks to extract the image features of size $8 \times 8 \times 2048$. These features are flattened (64×2048) and then passed on to a dense layer with number of units equal to the word embedding size, which is chosen to be 256. The outputs of this dense layer with shape 64×256 act as inputs to the RNN architecture in the first two models and to the Transformer in the third model, which acts as the decoder. The ground truths are passed to the decoder, where a tokenizer and an embedding layer are used to map the words to vectors in an embedding space. The tokenizer pads each sequence with a special token '<pad>' in order to ensure all the captions have the same length, that is equal to the length of the longest caption. Besides, special tokens '<start>' and '<end>' are used to indicate the start and the end of a caption. Each of these models then follows the architectures described in the previous section and have a dense SoftMax layer with 5,000 units (equal to the vocabulary size) as its output. All the implementations are performed using the TensorFlow deep learning library [23] for Python.

For training each of the three models, we use the Adam optimizer [24] with a constant learning rate of 0.001 and a batch size of 64. For the transformer model, the number of layers is limited to 4 as opposed to the base model in [18] which has 6. Besides, the number of attention heads is set to 4, i.e., the model employs four parallel attention layers. All the models aim to maximize the log likelihood of the reference caption given the training image.

4.3 Results

In Table 3, we compare the BLEU-1, 2, 3, 4 scores of our three models on the testing set with several other existing algorithms [1, 2, 21]. Google NIC [1] employs a pre-trained GoogLeNet [6] network for image extraction and uses an LSTM-based decoder for generating captions. The two models in [2] employ a pre-trained VGG-16 [10] network and an LSTM-based decoder, in addition to soft and

hard attention mechanisms. The final model, proposed by SR et al. in [21] employs a DenseNet [25] and a bi-directional LSTM-based decoder along with some game theoretic optimization. Figure 1 shows several images from the test set along with their captions. The images are grouped according to how good the captions are in describing the image contents.

Model	**BLEU Scores**			
	BLEU-1	**BLEU-2**	**BLEU-3**	**BLEU-4**
Google NIC, 2015 [1]	0.63	0.41	0.27	-
Soft-Attention, 2016 [2]	0.67	0.448	0.299	0.195
Hard-Attention, 2016 [2]	0.67	0.457	0.314	0.213
Game Theoretic Search, 2019 [21]	0.699	0.563	0.465	0.429
Our Model: M1	0.677	0.511	0.39	0.303
Our Model: M2	0.689	0.523	0.404	0.319
Our Model: M3	**0.706**	**0.539**	**0.43**	**0.365**

Table 3: BLEU-1, 2, 3, 4 scores of our models on Flick8k dataset compared to other algorithms

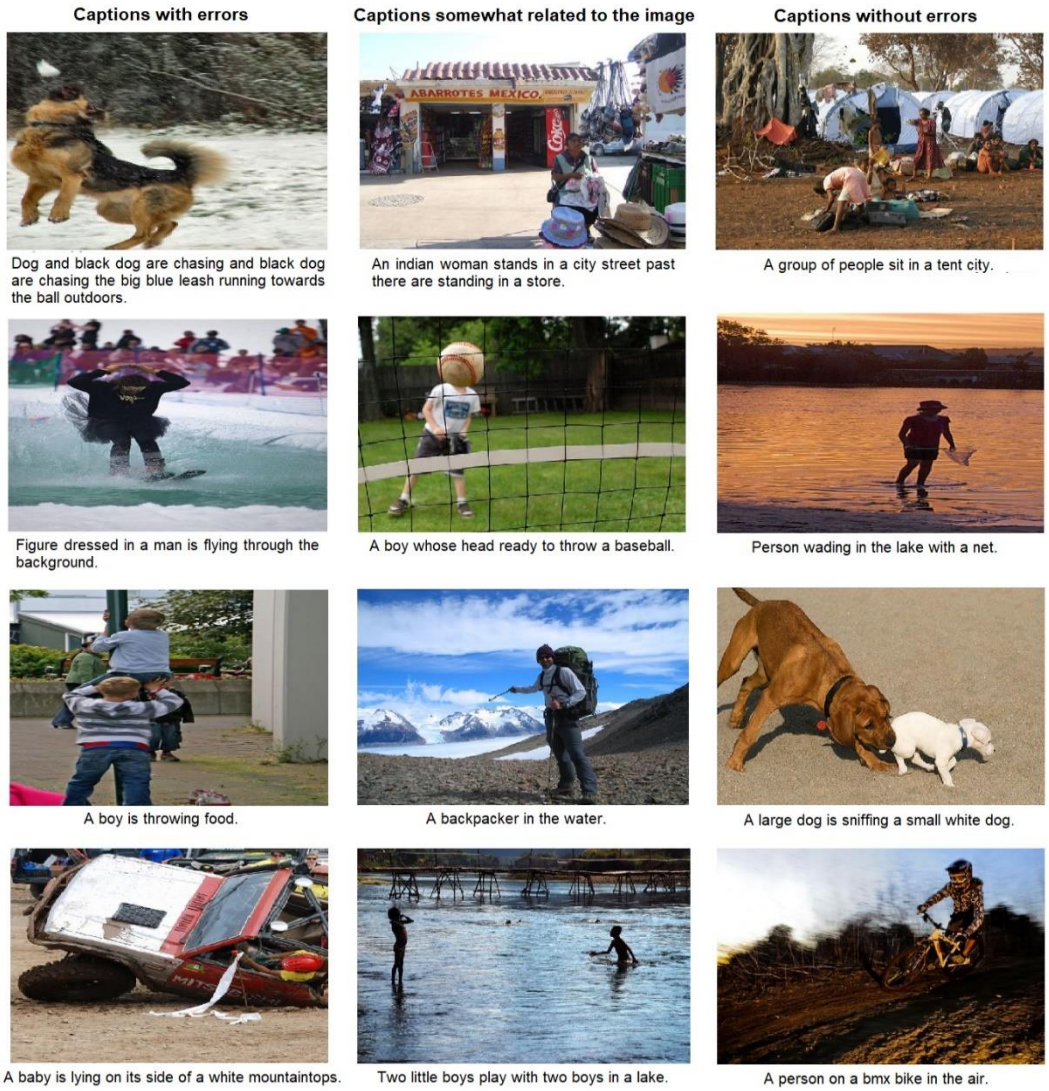
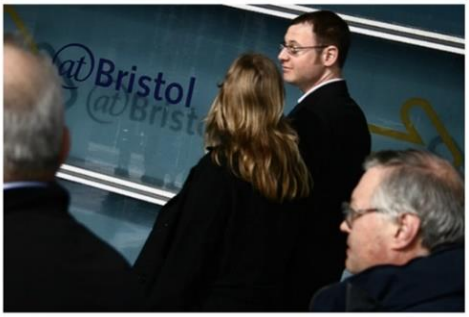


Figure 1: A few generated captions, grouped by their accuracy in describing the image

4.4 Analysis

We observed that replacing the GoogLeNet [6] architecture in the original model [1] with one of InceptionV3 [22] and Xception [19] networks, and replacing LSTM units with GRUs improved the BLEU scores. Even though we mainly attribute this improvement to the change in feature extractor, we also believe this could partly be due to the fact that GRUs can slightly outperform LSTM units when dealing with shorter sequences. In addition, we observed that there was a further increase in the scores upon replacing the decoder RNN with a Transformer [18]. We think this boost in the performance is primarily due to the presence of multiple attention heads in the model as it allows the model to jointly attend to information at different positions.



Reference Captions (Ground Truths)

- A group of adults are walking.
- A man in a black suit looking toward a woman with long hair.
- A man wearing glasses and a dark coat looks at something to the left of the picture.
- A woman talks to a man while walking in a crowd.
- Four people in dark clothes walk in the same direction.

Predicted Captions (BLEU-1 Score in [ ])

- M1: A group of men are talk while standing on one busy street while a group of men are walking in a crowd. [0.39]
- M2: A woman wearing a blue shirt and a man in a grey <UNK> is walking in a crowd. [0.53]
- M3: A group of people walking in a subway station. **[0.70]**

Reference Captions (Ground Truths)

- A man in roping a bull in a rodeo while others in cowboy hats watch.
- A man is up in the air with one hand on a bull.
- A rodeo rider gets tossed up into the air by a black bull as fellow cowboys look on.
- Man getting thrown in the air while bull riding.
- The man in black is flying off the full in front of a red fence.

Predicted Captions (BLEU-1 Score in [ ])

- M1: The girl is flying through the grass. [0.33]
- M2: A man is in <UNK> plays an charging bull. [0.50]
- M3: A man in black clothes is jumping off a red bull into a tire top. **[0.63]**

Reference Captions (Ground Truths)

- A man in a cap with a goatee leans against the window and reads a magazine.
- A man sitting on an empty train.
- A man with a black hat and black beard is reading a magazine sitting on a blue bench.
- A man with a hat sits in a booth and reads.
- The man reading a magazine is sitting in the large room.

Predicted Captions (BLEU-1 Score in [ ])

- M1: A man in a cap sitting on a blue bus. **[0.82]**
- M2: A man in a white hat and jeans sitting outside a window and leaning against the window and slide. [0.55]
- M3: A man with a goatee wears a black shirt and looking at magazine. [0.56]

Figure 2: Captions generated by our models on three images from test set (BLEU-1 scores in [])

Figure 2 depicts the reference captions and the captions generated by the three trained models on three images selected from our test set. Additionally, the BLEU-1 scores of these generated captions are also reported. The predictions made by our third model M3 on the first two images significantly outperform those made by the other two models both qualitatively and quantitatively. We credit this to the better recognition of key objects and their relationships. On the third image, we found that the caption generated by the first model has a much higher BLEU-1 score as compared with the other models even though each of the three models identifies a man and accurately describes only a part

of the image. The first model could identify that a man was wearing a cap and sitting on a blue bus. The second model, apart from identifying that the man wearing a hat and sitting, also identified that he was leaning against a window. The third model could identify that the man was sporting a goatee and looking at a magazine. Despite similar performances, the BLEU score of the first caption is much better than the other two. This is because of the overlap between the third reference and the first generated caption – the former contains 'sitting on a blue bench', the latter contains 'sitting on a blue bus'. The score is higher despite the wrong identification of object. We think this is one of the limitations of BLEU score in accurately assessing the quality of a generated caption.

5 Summary

Our simpler models which employed InceptionV3 + GRU and Xception + GRU were able to slightly exceed the benchmark performances of the original papers [1, 2] while simultaneously reducing the parameters size and computation cost. Besides, our transformer model was able to achieve better performance than the recurrent model counterparts with limited training proving once again that "Attention is All You Need" [18] to generate context rich sentences from images. Given the limitations of the hardware resources, our experiments involved only a few variants in the model parameters. We see a great scope for improvement on our current design and the possibility to employ a few changes such as making use of pre-trained word embedding schemes such as Word2Vec [28] or GloVe [29]. From generating small captions, we can extend to generating a meaningful description such as a summary of few sentences given an image if we can employ pre-trained text generation models such as BERT [14]. Our results also showcased some loopholes in the BLEU scoring scheme and presents the case for designing a better scheme.

Acknowledgement

We would like to thank our CSE 676 Deep Learning course instructors Dr. Changyou Chen, Le Fang and Zhenyi Wang for their insightful lectures and their immense contribution in igniting our interest in the subject. Additionally, we also thank our teaching assistant Mohammad Abuzar Shaikh for being helpful throughout the semester and providing guidance in this project.

References

[1] Vinyals, Oriol, et al. "Show and tell: A neural image caption generator." *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2015.

[2] Xu, Kelvin, et al. "Show, attend and tell: Neural image caption generation with visual attention." *International conference on machine learning*. 2015.

[3] Cho, Kyunghyun, et al. "Learning phrase representations using RNN encoder-decoder for statistical machine translation." *arXiv preprint arXiv:1406.1078* (2014).

[4] Bahdanau, Dzmitry, Kyunghyun Cho, and Yoshua Bengio. "Neural machine translation by jointly learning to align and translate." *arXiv preprint arXiv:1409.0473* (2014).

[5] Sutskever, Ilya, Oriol Vinyals, and Quoc V. Le. "Sequence to sequence learning with neural networks." *Advances in neural information processing systems*. 2014.

[6] Szegedy, Christian, et al. "Going deeper with convolutions." *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2015.

[7] Deng, Jia, et al. "Imagenet: A large-scale hierarchical image database." *2009 IEEE conference on computer vision and pattern recognition*. Ieee, 2009.

[8] Mao, Junhua, et al. "Explain images with multimodal recurrent neural networks." *arXiv preprint arXiv:1410.1090* (2014).

[9] Kiros, Ryan, Ruslan Salakhutdinov, and Rich Zemel. "Multimodal neural language models." *International conference on machine learning*. 2014.

[10] Simonyan, Karen, and Andrew Zisserman. "Very deep convolutional networks for large-scale image recognition." *arXiv preprint arXiv:1409.1556* (2014).

[11] Hossain, MD Zakir, et al. "A comprehensive survey of deep learning for image captioning." *ACM Computing Surveys (CSUR)* 51.6 (2019): 1-36.

[12] Lu, Jiasen, et al. "Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks." *Advances in Neural Information Processing Systems*. 2019.
[13] Qi, Di, et al. "Imagebert: Cross-modal pre-training with large-scale weak-supervised image-text data." *arXiv preprint arXiv:2001.07966* (2020).
[14] Devlin, Jacob, et al. "Bert: Pre-training of deep bidirectional transformers for language understanding." *arXiv preprint arXiv:1810.04805* (2018).
[15] Sharma, Piyush, et al. "Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning." *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. 2018.
[16] Chen, Xinlei, et al. "Microsoft coco captions: Data collection and evaluation server." *arXiv preprint arXiv:1504.00325* (2015).
[17] Plummer, Bryan A., et al. "Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models." *Proceedings of the IEEE international conference on computer vision*. 2015.
[18] Vaswani, Ashish, et al. "Attention is all you need." *Advances in neural information processing systems*. 2017.
[19] Chollet, François. "Xception: Deep learning with depthwise separable convolutions." *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2017.
[20] Papineni, Kishore, et al. "BLEU: a method for automatic evaluation of machine translation." *Proceedings of the 40th annual meeting on association for computational linguistics*. Association for Computational Linguistics, 2002
[21] SR, Sreela, and Sumam Mary Idicula. "Dense Model for Automatic Image Description Generation with Game Theoretic Optimization." *Information* 10.11 (2019): 354.
[22] Szegedy, Christian, et al. "Rethinking the inception architecture for computer vision." *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2016.
[23] Abadi, Martín, et al. "Tensorflow: A system for large-scale machine learning." 12th {USENIX} Symposium on Operating Systems Design and Implementation ({OSDI} 16). 2016.
[24] Kingma, Diederik P., and Jimmy Ba. "Adam: A method for stochastic optimization." *arXiv preprint arXiv:1412.6980* (2014).
[25] Huang, Gao, et al. "Densely connected convolutional networks." *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2017.
[26] Cho, Kyunghyun, et al. "Learning phrase representations using RNN encoder-decoder for statistical machine translation." *arXiv preprint arXiv:1406.1078* (2014).
[27] Hochreiter, Sepp, and Jürgen Schmidhuber. "Long short-term memory." *Neural computation* 9.8 (1997): 1735-1780.
[28] Mikolov, Tomas, et al. "Efficient estimation of word representations in vector space." *arXiv preprint arXiv:1301.3781* (2013).
[29] Pennington, Jeffrey, Richard Socher, and Christopher D. Manning. "Glove: Global vectors for word representation." *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*. 2014.